Electronic Version
Stylesheet Version v1.1.1

Description

[FLASH MEMORY WITH SELF-ALIGNED SPLIT GATE AND METHODS FOR FABRICATING AND FOR OPERATING THE SAME]

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application is a divisional of a prior application serial no. 10/249,024, filed March 11, 2003.

BACKGROUND OF INVENTION

[0002] Field of the Invention

[0003] The present invention relates to a non-volatile memory (NVM) and the methods for fabricating and for operating the same. More particularly, the present invention relates to a flash memory with a self-aligned spilt gate and methods for fabricating and operating the same.

[0004] Background of the Invention

[0005] Flash memory can retain information even when power is

interrupted and is small in size, faster in reading/programming and can resist vibration, so it is widely used. A flash memory comprises a floating gate and a control gate that are isolated by a dielectric layer, wherein the floating gate is isolated from the substrate by a tunnel oxide layer. During the writing/erase operation, electrons are injected into/ejected from the floating gate with a voltage applied to the control gate. During the reading operation, a working voltage is applied to the control gate. At this time, the charging state on the floating gate causes a conducting status of ON or OFF of the channel that is under the floating gate. The conducting state of ON/Off corresponds to the data of 0/1The data in the above mentioned flash memory is erased by increasing the potential of the substrate, the drain/source or the control gate relative to the floating gate. The electrons ejected from the floating gate flow into the substrate or the drain/source via the tunnel oxide layer by tunneling. This mechanism is known as substrate erase mechanism or drain/source side erase mechanism. Another mechanism is to eject the electrons in the floating gate to the control gate via the dielectric layer. However, the amount of the electrons ejected from the floating gate is difficult

to precisely control during erasing. If too many electrons are ejected from the floating gate, the floating gate has net positive charges. This phenomenon is called "over-erasing". When the over-erasing effect is severe, the channel under the floating gate is switched on even when the working voltage is not applied to the control gate. This may lead to an error in data reading. Therefore, a split gate design is adopted in many kinds of flash memory. One of the characteristics of the split gate is that the control gate has a portion above the floating gate and another portion above the substrate with separation of a gate dielectric layer. Thus when the over-erasing occurs to switch on the channel under the floating gate even if there is no working voltage applied, the channel under the control gate remains closed. Therefore, the drain and the source still cannot be electrically connected. This prevents the data from being erratically determined.

[0006] The process for fabricating the split gate flash memory in the prior art is described as following with reference to FIG. 1A to 1D.

[0007] As shown in FIG. 1A, a substrate 100 is provided. A gated oxide 104, a polysilicon layer 106 and a dielectric layer 108 is sequentially formed on the substrate 100, wherein

the polysilicon layer 106 serves as a floating gate. A thermal oxidation process is carried out to form an oxide layer 110 on the sidewalls of the polysilicon layer 106 and on the substrate 100.

[0008] As shown in FIG. 1B, a conformal polysilicon layer 112 is formed on the substrate 200 covering the dielectric layer 108 and oxide layer 110.

[0009] As shown in FIG. 1C, a photolithography process and an etching process are performed to form control gates 112a and 112b covering a portion of floating gate 106 and a portion of the substrate 100. An ion implantation process is carried out to form a common source 116 in the substrate 100 between the control gates 112a and 112b and a drain 114 in the substrate 100 on the other side of the floating gate 106.

[0010] There are some problems in the conventional method for fabricating the spilt gate flash memory. One is that the control gates 112a and 112b have non-uniform width as shown in FIG. 1D. Since the patterning process is not carried out by using a self-aligned method, the misalignment of the photolithography process will lead to asymmetric control gates 112a and 112b. Therefore, the size of the control gate, the channel length and the channel current

of each memory are also not constant. This affects the quality of the product. The other problem is that the process window in the conventional method is small since a self-aligned method is not used. This causes a disadvantage that the cell dimension is hard to scale down. Another problem is that two adjacent memory cells are unsymmetrical and have different electric properties since the memory cells are not formed on strip-like active regions.

SUMMARY OF INVENTION

[0011] The present invention provides a flash memory with a self-aligned spilt gate and the methods for fabricating and for operating the same to solve the problem that the adjacent memory cells are unsymmetrical and not potential equivalent as in the prior art.

[0012] The present invention also provides a method for operating a flash memory with a self-aligned spilt gate in order to reduce the operating voltage in the programming, erasing or reading operation.

[0013] The present invention provides a flash memory with a self-aligned spilt gate. The flash cell consists of a substrate, a deep n-type well and a shallow p-type well, a gate oxide layer, a control gate, a capping layer, a floating

gate, a tunnel oxide layer, a drain, a common source and a pocket p-well. The deep n-type well is located in the substrate and the shallow p-type well is located in the deep n-well. The control gate is located on the substrate covering a portion of the shallow p-type well. The gate oxide is located between the control gate and the substrate and the capping layer is located on the control gate. The floating gate is located on one sidewall of the control gate and the capping layer and over the substrate. The tunnel oxide layer is located between the control gate and the floating gate and between the floating gate and the substrate. In the present invention, a dielectric spacer is further on the other sidewall of the control gate and the capping layer to protect the control gate from being damaged during a subsequent metal interconnection process. The drain is located in the deep n-type well under the dielectric spacer and adjacent to the control gate. A common source is located in and connected to the deep n-type well and under extending to a portion of the floating gate and adjacent to the shallow p-type well. A pocket p-well is located in the substrate around the drain and electrically connecting the divided shallow p-type wells beside the drain to make the shallow p-type well of each cell po-

tential equivalent.

[0014] The present invention provides a method for fabricating a flash memory with a self-aligned spilt gate. An isolation is formed on a substrate to define an active region. A deep n-type well is formed in a substrate and a shallow p-type well is formed in the deep n-well. A gate oxide layer, a control gate and a capping layer are formed on a portion of the shallow p-type well. A tunnel oxide layer is formed on the sidewalls of the control gate and on the substrate by conducting a thermal process. A conformal conductive layer is formed covering the capping layer and the substrate. The conformal conductive layer is etched back to form a conductive spacer on the sidewalls of the capping layer and the control gate. Thereafter, the conductive spacer on one side of the control gate is removed to leave the conductive spacer on the other side as a floating gate. A common source is formed in and connects to the deep n-well, wherein the common source under extends to a portion of the floating gate about a half of the floating gate width. A drain is formed in the shallow p-type well adjacent to the control gate. A dielectric spacer is formed on the sidewall of the control gate without the conductive spacer formed thereon to protect the control gate from

being damaged in subsequent etching processes. A pocket p-well is formed around the drain to connect with the shallow p-type well beside the drain.

[0015] The present invention provides a method for operating a spilt gate flash memory. The split gate flash memory cell comprises a substrate, a deep n-type well in the substrate, a shallow p-type well in the deep n-well, a gate oxide layer on the shallow p-type well , a control gate on the gate oxide layer, a capping layer on the control gate, a floating gate on one sidewall of the control gate and the capping layer and over a portion of the substrate, a tunnel oxide layer between the control gate and the floating gate and between the floating gate and the substrate, a dielectric spacer on the other sidewall of the capping layer and the control gate, a drain in the shallow p-type well and adjacent to the control gate, a common source located in and connected to the deep n-type well and under, extending to a portion of the floating gate about a half of the floating gate width, and a pocket p-well located in the substrate around the drain and electrically connecting with the shallow p-type well. During the programming, a first voltage, such as 2 volts, is applied to the control gate to turn it on. A second voltage, such as 10 volts, is ap-

plied to the common source, and the drain and the pocket p-well are ground. Since the common source and the whole deep n-type well have the second voltage (e.g., 10 volts), the floating gate is coupled with a voltage about one half of the second voltage (e.g., about 5 to 6 volts). Since the channel length is very short, a large electric field is established in the vertical direction and in the lateral direction of the substrate. Consequently, hot electrons are formed and injected into the floating gate through the tunnel oxide. Therefore, the split gate cell uses source side injection for programming. During the erase operation, a third voltage, such as 20 volts, is applied to the control gate to eject electrons from the floating gate to the control gate by Fowler-Nordheim tunneling. In another erase method, a positive voltage, such as 12 volts, is applied to the control gate, a negative voltage, such as -8 volts, is applied to the common source, and the drain and the pocket p-well are floated to eject electrons by Fowler-Nordheim tunneling. During the reading operation, V_{cc} is applied to the control gate, a fourth voltage, such as 1.5 volts, is applied to the drain, and the common source and the pocket p-well are ground.

[0016] The design of the present invention is that the common

source is picked up by the deep n-well, so the cell current unsymmetry problem caused by block-like active regions can be avoided. Furthermore, since the floating gate is self-aligned to the control gate, the misalignment of the two can be avoided.

[0017] Moreover, the flash memory cell of the present invention uses a stacked gate structure including only a floating and a control gate, so its fabrication is simpler than that of the flash memory in the prior art that uses a stacked gate structure consisting of three polysilicon layers.

[0018] Moreover, this flash memory is erased by ejecting the electrons from the floating gate to the control gate by Fowler-Nordheim tunneling, so the over-erase problem can be overcome. Therefore, the biases applied in programming, erase and reading operations of the split gate flash memory of the present invention is lower than those applied in the prior art.

[0019] It is to be understood that both the foregoing general description and the following detailed description are exemplary, and are intended to provide further explanation of the invention as claimed.

Brief Description of Drawings

[0020] The accompanying drawings are included to provide a

further understanding of the invention, and are incorporated in and constitute a part of this specification. The drawings illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention.

[0021] FIG.1A to 1D schematically illustrate the process flow of a split gate flash memory in the prior art in a cross-sectional view.

[0022] FIG. 2 schematically illustrates the top view of the self-aligned split gate flash memory according to a preferred embodiment of the present invention.

[0023] FIG. 3A to 3H schematically illustrate the process flow of fabricating the spilt gate flash memory in FIG. 2 in a cross-sectional view along line I-I".

[0024] FIG. 4 illustrates a circuit diagram of the self-aligned split gate flash memory according to the preferred embodiment of the present invention.

[0025] FIG. 5 shows a flow chart of fabricating the self-aligned split gate flash memory according to the preferred embodiment of the present invention.

[0026] FIG. 6 schematically illustrates the top view of the self-aligned split gate flash memory after the common source is formed according to the preferred embodiment of the

present invention.

DETAILED DESCRIPTION

[0027] FIG. 2 schematically illustrates the top view of the self-aligned split gate flash memory according to the preferred embodiment of the present invention, whereintwo pairs of memory cells are shown in each row of the memory array. FIG. 3A to 3H schematically illustrate the process flow of fabricating the spilt gate flash memory in FIG. 2 in a cross-sectional view along line I-I". FIG. 5 shows a flow chart of fabricating the self-aligned split gate flash memory according to the preferred embodiment of the present invention.

[0028] Referring to FIG. 2, FIG. 3A and FIG. 5, a substrate 200, such as a p-type silicon substrate, is provided. Isolation (non shown) is formed on the substrate to define an active region 201. A deep well 202, such as an n-type deep well, is formed in the substrate 200 (step 500). A shallow well 204, such as a p-type shallow well, is then formed in the deep well 202 by, for example, an ion implantation process (step 502). The implanted ion used in the ion implantation process for forming the p-type shallow well includes boron ion. The implanting energy in the ion implantation process is about 20KeV and the dosage is

about $1x10^{12}/cm^2$.

[0029] Thereafter, a gate oxide layer 206 is formed on the substrate 200 (step 504). The gate oxide layer 206 is formed with a thermal process, for example, and is about 250 Angstroms thick. A conductive layer 208 is formed on the gate oxide layer 206. The conducting layer 208 is, for example, a polysilicon layer with a thickness of 600 Angstroms and the polysilicon is doped for reducing the resistance. The method of doping the polysilicon layer comprises an ion implantation process. The implanted ion includes arsenic ion. The implanting energy in the ion implantation process is about 30KeV and the dosage is about $1x10^{15}/cm^2$. A capping layer 210, such as an oxide layer with a thickness of about 3500 Angstroms, is formed on the conducting layer 208.

[0030] Refering to FIG. 2, 3B and FIG. 5, a photolithography process and an etching process are conduced for patterning the capping layer 210 and the conductive layer 208 to form patterned capping layers 210a and 210b and two control gates 208a and 208b (step 506). In the patterning method, for example, a patterned resist is formed on the capping layer 210 and then the capping layer 210 is etched to form the patterned capping layers 210a and

210b with the patterned resist as a mask. The patterned resist is removed and the conductive layer 208 is etched to form the control gates 208a and 208b with the patterned capping layer 210a and 210b as a mask layer. Thereafter, the gate oxide layer 206 exposed by the control gate 208a and 208b is removed with a cleaning process to expose the substrate 200, while the gate oxide layer 206a and 206b under the control gate 208a and 208b are left.

[0031] As shown in FIG. 3C and FIG. 5, a thermal process is performed to form a tunnel oxide layer 212 on the sidewalls of the control gate 208a and 208b and on the substrate 200 (step 508). Since the control gates 208a and 208b are doped with arsenic ions, the tunnel oxide layer 212 formed on the sidewalls of the control gates 208a and 208b is thicker than the tunnel oxide layer 212 formed on the substrate 200. For example, the thickness of the former is about 300 Angstroms and that of the latter is about 90 Angstroms.

[0032] Thereafter, a conformal conductive layer 214 is formed on the substrate 200 covering the patterned capping layer 210a and the tunnel oxide layer 212. The conformal conducting layer 214 is, for example, a polysilicon layer

doped with phosphorous ions and with a thickness of about 3000 Angstroms. A method for doping the polysilicon layer comprises, for example, conducting an implantation process to introduce phosphorous ions into the polysilicon layer, wherein the implanting energy is about 60 KeV and the dosage is about $5x10^{14}/cm^2$.

[0033] As shown in FIG. 3D, the conformal conducting layer 214 is etched back to form a conductive spacer 214a on the sidewalls of the control gate 208a and 208b and the patterned capping layer 210a. The tunnel oxide layer 212 isolates the conductive spacer 214a from the control gate 208a and also isolates the conductive spacer 214a from the substrate 200.

[0034] As shown in FIG. 2, 3E and 5, a photolithography process and an etching process are conducted to remove a portion of the conductive spacer 214a so that a pair of conductive spacers 214a and 214b are left on two sides of the control gate pair 208a and 208b on the active region 201 (step 510). The conductive spacers 214a and 214b left on the sidewall of the control gate 208a and 208b serve as a floating gate.

[0035] The floating gate of the split gate flash memory of the present invention is formed by etching back the conformal

conducting layer 214, so the floating gate is self-aligned to the control gate 208a. Therefore, misalignment and the problems arising therefrom can be avoided.

[0036] Thereafter, two common sources 218a and 218b are formed in the deep well 202 through the shallow well 204 adjacent to the floating gates 214a and 214b, respectively (step 512). The common sources 218a and 218b are formed by performing an ion implantation process with a mask layer 250 shown in FIG. 6, wherein a region surrounded by the mask layer 250 is implanted with ions. The implanted ion is, for example, an n-type ion. The implanting energy is about 60 KeV and the dosage is about $1x10^{14}/cm^2$. A thermal process is conducted at, for example, $600^oC \sim 900^oC$ in order to drive the dopants into the substrate 200. Therefore, the common source 218a and 218b extends to the substrate under a portion of the floating gate 214a and 214b and connects with the deep well 202. The portion of the common source 218a and 218b under the floating gate 214a and 218b has a width about a half of the width of the floating gate 214a and 214b. Another photolithography process and another ion implantation process are performed to form a drain 216 such as n-type drain in the shallow well 204 between the

control gate 208a and 208b (step 514). The dosage for forming the drain 216 is higher than that for forming the common source 218.

[0037] As shown in FIG. 3F and 5, a pair of dielectric spacers 220a and 220b are formed on the sidewalls of the control gate 208a and 208b where the floating gate 214a and 214b are not formed (step 516). Thus, the control gate 208a and 208b can be protected from damages during the subsequent etching process of interconnect. The material of the dielectric spacer 220a and 220b is, for example, silicon nitride or silicon oxide.

[0038] As shown in FIG. 2, 3G and 5, a pocket ion implantation process is conducted to form a pocket well 224, such as an pocket p-well, around the drain 216 and into the deep well 202 through the shallow well 204 (step 518). The dosage of the pocket ion implantation process is about $1x10^{13}/cm^2$, and the implanting energy is enough for penetrating the isolation. Since the thickness of the control gate 208a and the capping layer 210a is about 4000 Angstroms, the implanted ions can not penetrate the control gate 208a. The purpose of forming the pocket well 224 is to connect the shallow wells 204 in the same column of memory cell pairs to make all of the shallow wells

204 connected each other.

[0039] Refer to FIG. 3H, an interlayer dielectric layer (ILD) 226 is formed on the substrate 200, and then a contact 230 and wiring line 228 are formed in and on the interlayer dielectric layer 226, respectively. Moreover, in the preferred embodiment of this invention, other contacts 240 can be formed to pick up the pocket p-wells 224 to connect the pocket p-wells 224 with other conductive structures, as shown in FIG. 2.

[0040] As shown in FIG. 3H, the self-aligned split gate flash cell of the preferred embodiment consists of a substrate 200 having a deep n-type well 202 and a shallow p-type well 204, a gate oxide layer 206a, a control gate 208a, a capping layer 210a, a floating gate 214a, a tunnel oxide layer 212, a dielectric spacer 220, a drain 216, a common source 218a and a p-type pocket well 224.

[0041] The shallow p-type well 204 is located in the deep n-type well 202. The control gate 208a is located on the gate oxide layer 206a formed on the p-shallow well 204. The capping layer is located on the control gate 208a. The floating gate 214a is located on one sidewall of the control gate 208a and the capping layer 210a and over the substrate 200. The tunnel oxide layer 212 is located be-

tween the control gate 208a and the floating gate 214a and between the floating gate 214a and the substrate200. The dielectric spacer 220 is located on the other sidewall of the control gate 208a and the capping layer 210a. The drain 216 is located in the substrate 200 under the dielectric spacer 220 adjacent to the control gate 208a. The common source 218a is located in the substrate 200 adjacent to the shallow p-type well 204 and into the deep well 202, and extends to the substrate 200 under a portion of the floating gate 214a. The p-type pocket p-well 224 is located in the substrate 200 around the drain 216 and serves to electrically connect the divided shallow p-type well s 204 beside the drain 216.

[0042] The difference between the split gate flash cells of the present invention and the prior art lies in the fact that the relative position of the control gate and floating gate is reversed. Because of the aforementioned structure, the performance of the split gate flash memory of the present invention is better than that in the prior art. The reasons are described as follows.

[0043] Refer to FIG. 4, FIG. 4 illustrates a circuit diagram of the self-aligned split gate flash memory according the preferred embodiment of the present invention.

[0044] As shown in FIG. 4, the split gate of the present invention comprises a cell array, the word lines WL and the bit lines BL perpendicular to the word lines WL, wherein each cell is selected by one word line WL and one bit line BL.

[0045] In each cell, the drain is electrically coupled to a bit line BL and the control gate is electrically coupled to a word line WL. The sources are electrically coupled to each other by the deep n-well, so a common source CS is formed with equal potential. Therefore, in a memory block, all memory cells share a common source. The pocket p-well is used to connect the shallow p-type well s in the same column of memory cell pairs. Refer to FIG. 4, since the shallow p-type well s of the cells in the same column are picked up by the pocket p-wells, the shallow p-type well of each cell is potential equivalent.

[0046] A method of operating a split gate flash cell of the present invention is described as follow: This split gate cell uses source side injection in programming. During the programming, a first voltage, such as 2 volts, is applied to the control gate to turn it on. A second voltage, such as 10 volts, is applied to a common source, and the drain and the pocket p-well are grounded. Because the common source and the whole deep n-type well are in the second

voltage (e.g., 10 volts), the voltage coupled to the floating gate is about one half of the second voltage (e.g., about 5 to 6 volts). Since the channel length under the floating gate is very short and the thickness of the tunnel oxide on the sidewall of the control gate is about 300 Angstroms, a large electric field is established in the vertical direction and the lateral direction of the substrate. Consequently, hot electrons are formed there and are injected into the floating gate through the tunnel oxide.

[0047] During an erase operation, a third voltage, such as 20 volts, is applied to the control gate and the common source, the drain and the pocket p-well are all floated to eject electrons from the floating gate to the control gate by Fowler-Nordheim tunneling. In another erase method, a positive voltage, such as 12 volts, is applied to the control gate, a negative voltage, such as -8 volts, is applied to the common source, and the drain and pocket p-well are floated to erase the data by Fowler-Nordheim tunneling.

[0048] During a reading operation, V_{cc} is applied to the control gate, a forth voltage, such as 1.5 volts, is applied to the drain, and the common source and the pocket p-well are ground.

[0049] Moreover, the design of the present invention is that the common source is picked up by the deep n-well, so the cell current unsymmetry problem caused by block-like active regions does not occur. Furthermore, since the floating gate is self-aligned to the control gate, the misalignment of the two can be avoided.

[0050] Moreover, the flash memory cell of the present invention uses a stacked gate structure including only a floating and a control gate, so its fabrication is simpler than that of the flash memory in the prior art that uses a stacked gate structure consisting of three polysilicon layers. Furthermore, since this flash memory cell is programmed by source side injection, a low voltage is sufficient for the programming.

[0051] Moreover, this flash memory is erased by ejecting the electrons from the floating gate to the control gate by Fowler-Nordheim tunneling, so the over-erase problem can be overcome. Therefore, the biases applied in programming, erase and reading operations of the split gate flash memory of the present invention is lower than those applied in the prior art.

[0052] It will be apparent to those skilled in the art that various modifications and variations can be made to the structure

of the present invention without departing from the scope or spirit of the invention. In view of the foregoing, it is intended that the present invention covers modifications and variations of this invention provided they fall within the scope of the following claims and their equivalents.